National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PDS4 – Some Principles for Agile Data Curation

AGU 2015 - IN22A-04
15 December 2015

John Steven Hughes, Daniel J Crichton, Sean H Hardman,
Ronald Joyner, Stirling Algermissen and Jordan Padams

Jet Propulsion Laboratory
California Institute of Technology

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Introduction

- Agile Data Curation
- Architecture
- Information Model
- Information Model-Driven Architecture

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Agile Data Curation

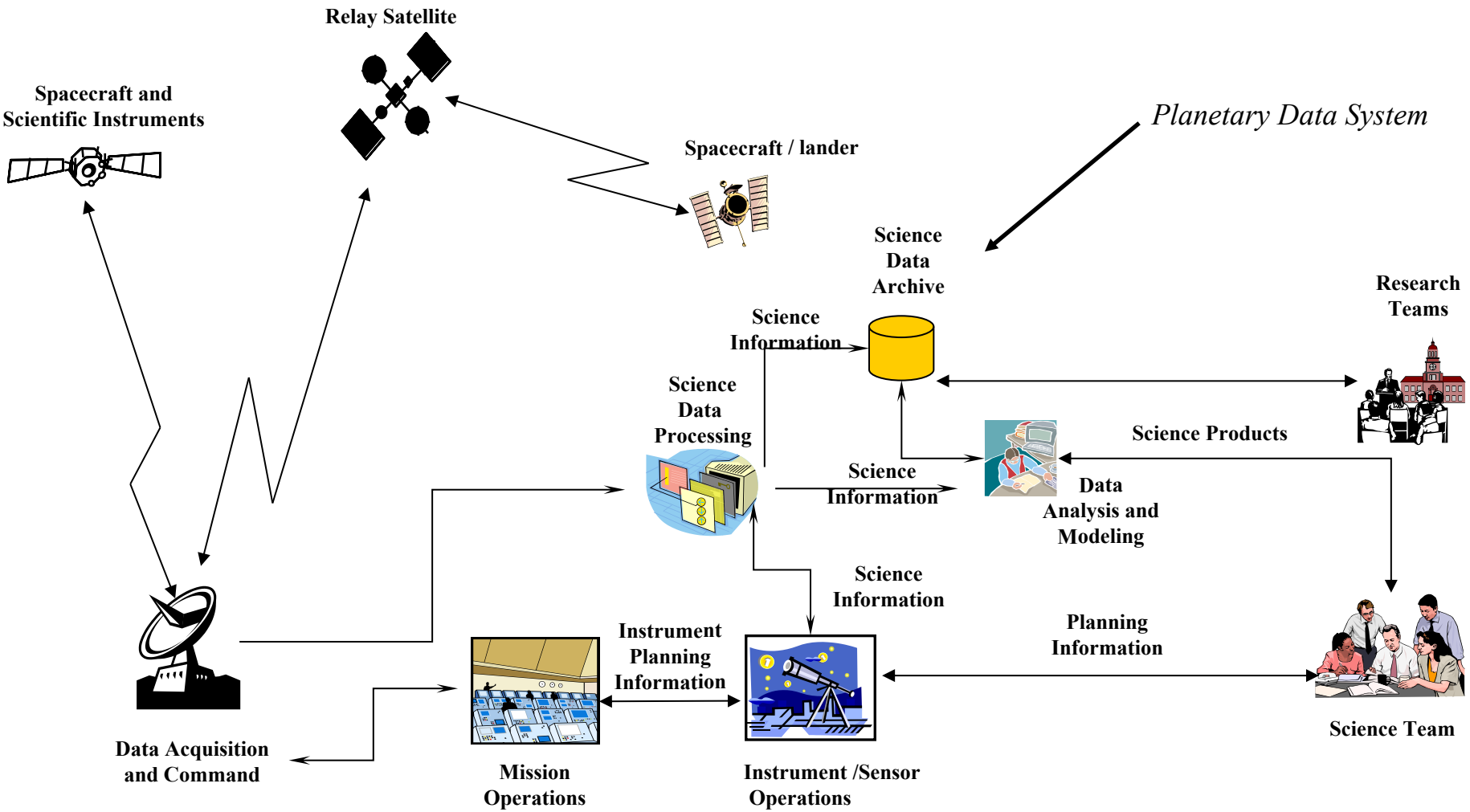
- PDS4, a research data management and curation system for NASA's Planetary Science Archive, was developed with agile data curation[1] principles in mind.
  - *adaptive planning*
  - *early delivery*
  - *evolutionary development*
  - *continuous improvement*
  - *rapid and flexible response to change*
- Using these principles an architectural process was used to plan, design, and construct a system that met both the desired functional and technical requirements and the community's expectations.
- The resulting architecture
  - *produces better research data products*
  - *uses less resources (time, effort, and money)*
  - *maximizes product usefulness for current and future scientists.*

[1] Young et al., Taking Another Look at the Data Management Life Cycle: Deconstruction, Agile, and Community, AGU 2014

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Reference Architecture

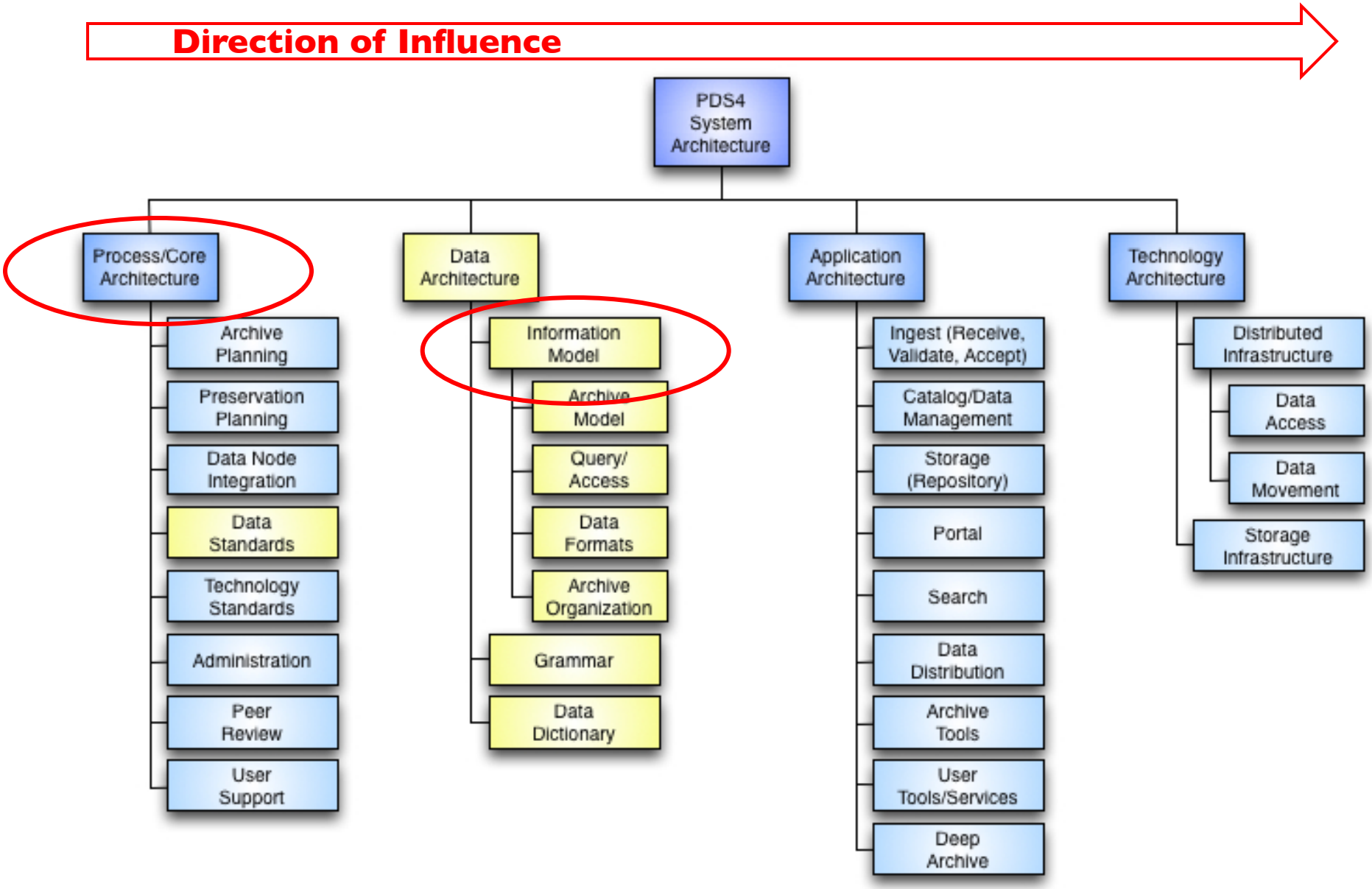


Credit: CCSDS Reference Architecture for Space Information Management

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

System Architecture



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Information Model - Definition

- "An information model is a representation of concepts, relationships, constraints, rules, and operations to specify data semantics for a chosen domain of discourse." [1]

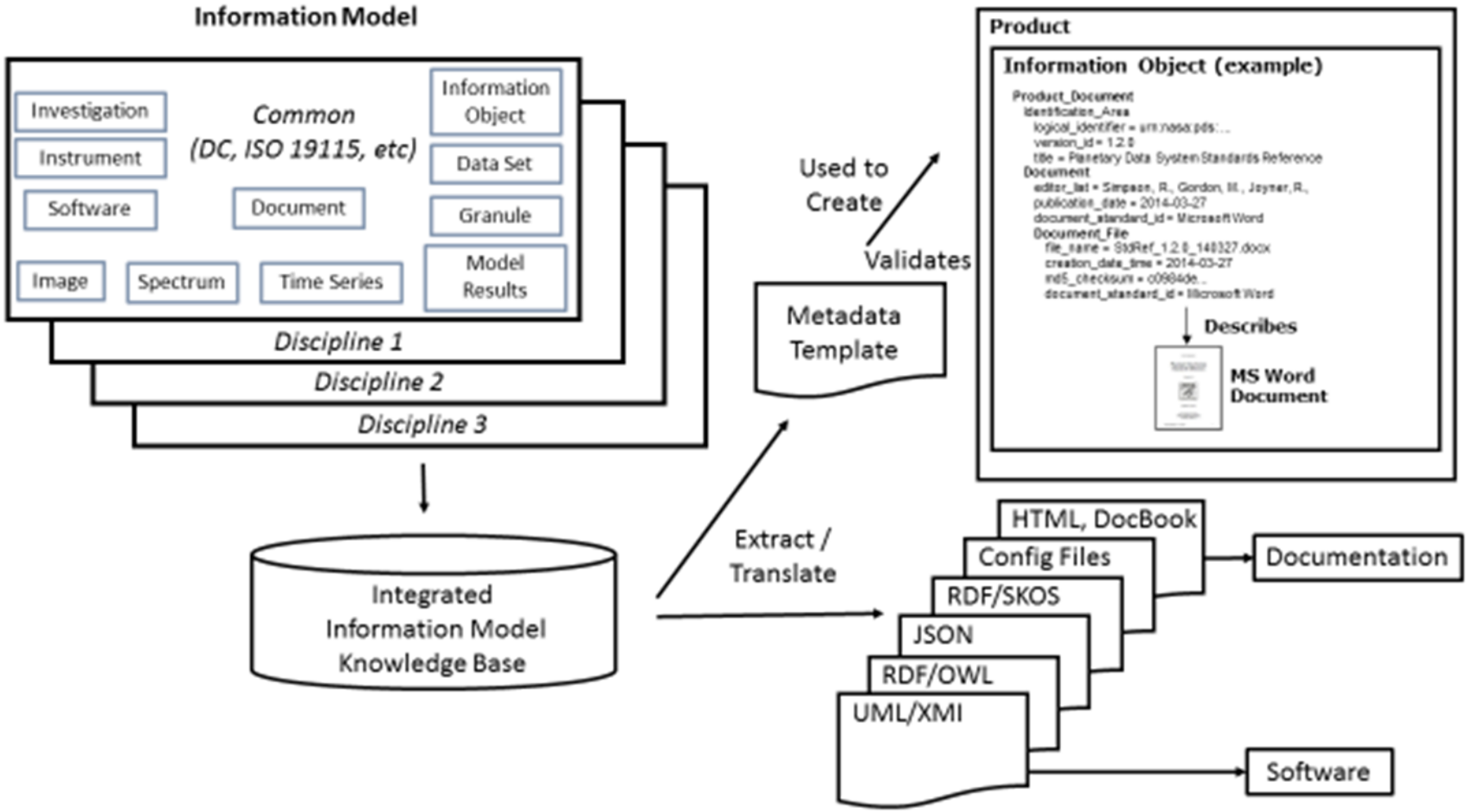
- It provides a sharable, stable, and organized structure of information requirements or knowledge for the domain context.

[1] Lee, Y.T. 1999. Information Modeling: From Design To Implementation. In Proceedings of the Second World Manufacturing Congress, ed. S. Nahavandi and M. Saadat, 315-321. Canada/Switzerland: International Computer Science Conventions.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Information Model in Context



Information Modeling Principles

- The Information Model is developed and maintained independent of the infrastructure's process, application and technology architectures.
  - *The information model needs to evolve independent of the information technology*
  - *Disentangles the model from the implementation*
  - *Provides a sharable, stable, and formal set of data requirements*
- The information architecture is based on an ontology-based information model developed to leverage best practices from standard reference models for digital archives, digital object registries, and metadata registries
- Captures domain knowledge from a panel of planetary science domain experts.
  - *The modeling language should be formal, standardized, and broadly used and accepted*
  - *The language should be more expressive than the other languages in the system's framework.*

Purpose of the Information Model

- Defines the data structure (format)
- Defines the interpretation of the data
- Defines the context within which the data was captured, processed, and archived
- Defines the relationships between the data

National Aeronautics and Space Administration

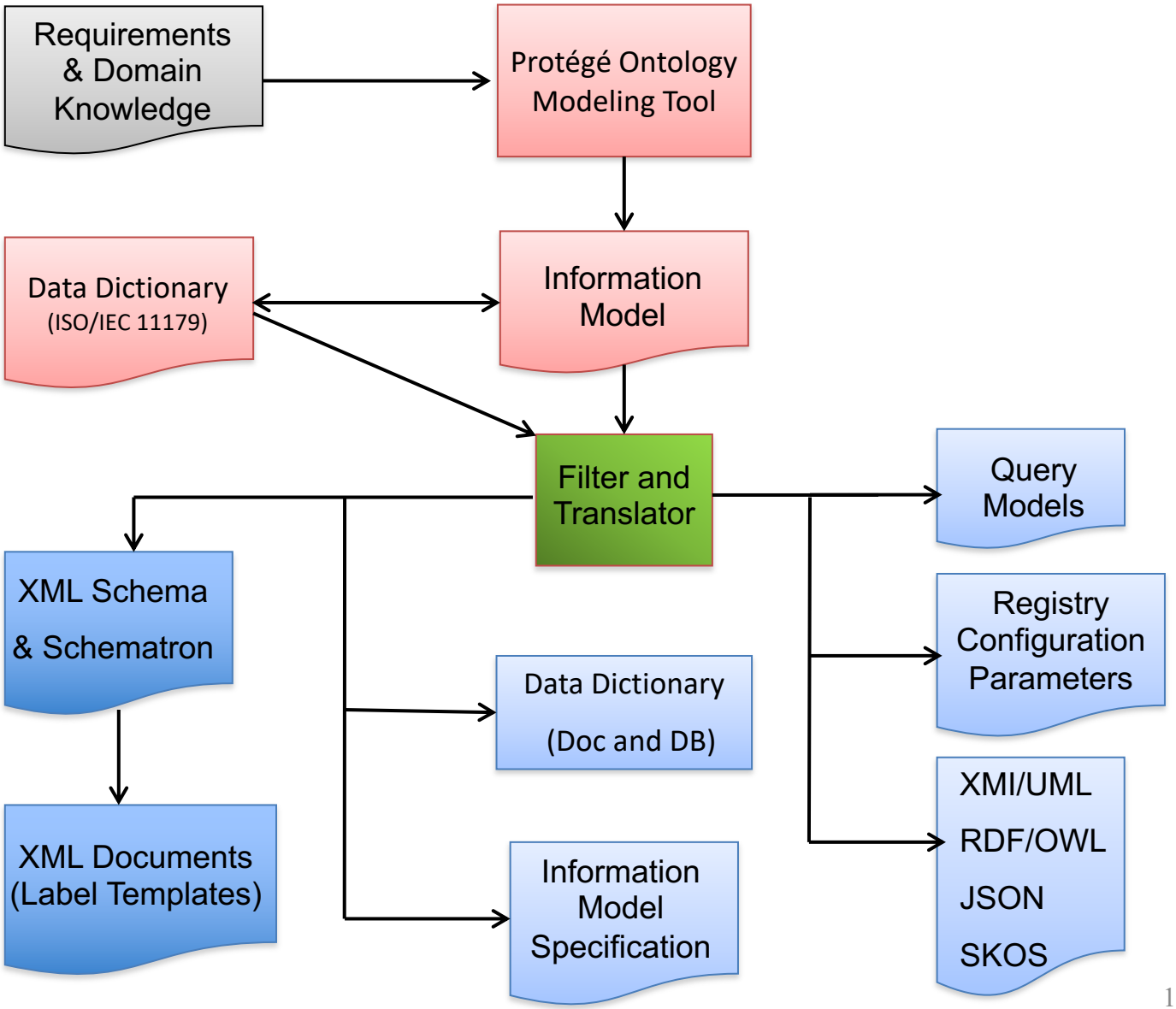
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Life-Cycle of the Information Model

- Define the scope and purpose of the model.
- Institute governance over the elements of the model
  - *Decide on the number of levels of control required*
  - *Allow controlled evolution at all levels.*
- Acquire domain knowledge from discipline experts
  - *Capture the knowledge in a modeling tool*
  - *Formalize the knowledge in a commonly acceptable and highly expressive language*
  - *Test early and frequently to determine the applicability and desired level of detail.*
- Use the model anytime and anywhere the knowledge is applicable
  - *Filter out the appropriate aspects of the model*
  - *Translate the filtered results to the most appropriate language for the purpose.*

1..*

Knowledge Data Flow



PDS4 Agile Curation

- Adaptive planning
  - Software and services are designed to respond to the information model.
  - An extensible model was designed to accommodate new features.
  - Controlled adaptation occurs at all levels of governance but most easily at the mission/project level
  - Knowledge acquisition keeps the domain experts “in-the-loop”.

- Evolutionary development
  - Prior to V1.0, versions of the model were frequently released to allow users to test and respond.
  - After V1.0, new versions of the model are released regularly with Change Control Board (CCB) approved changes.
  - The information, process, and technology components of the architecture evolve independently and at different speeds.

PDS4 Agile Curation

- Early delivery
  - First drafts of design elements were/are generated upon request for review and comment.
  - Versions of the model were/are made available upon request for review and testing.
  - Software and services were/are built, configured, and deployed regularly for review and testing.

- Continuous improvement
  - Stable versions of the model are released with regularly scheduled system builds.
  - The change control process in concert with the multi-level governance structure is used to manage change.

Agile Curation in Practice

- Rapid and flexible response to change
  - *Changes at the mission/project level is localized to allow rapid and flexible response.*
  - *Multi-level governance provides appropriate controls.*
  - *Independent information, process, and technology architectures allow localized changes.*

Summary

- PDS4, a research data management and curation system for NASA's Planetary Science Archive, was developed with agile data curation principles in mind.

- The Information System's architecture allows independent development and evolution of the information, application, and technology components.

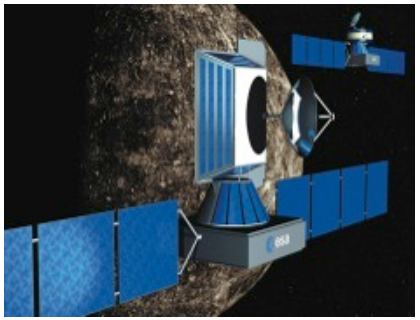
- A model-driven architecture, extensive involvement of domain experts, early delivery for review and testing, and continuous improvement have resulted in a set of standards, services, and software now being adopted by both the NASA and International planetary science community.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

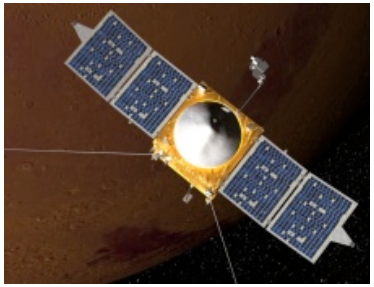
Adoption of PDS4

InSight (NASA)

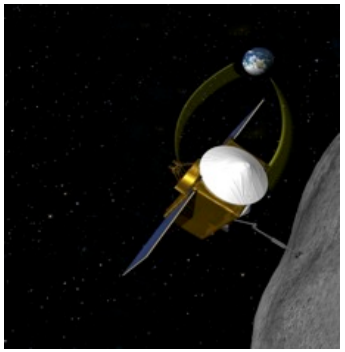


BepiColumbo (ESA/JAXA)

Planetary Data System Version 4
International, distributed, model-driven data architecture for capturing, managing and distributing planetary science data results to the world-wide science community.*
2000: 4 TBs; 2014: 720 TBs



MAVEN (NASA)



Osiris-REx (NASA)

LADEE (NASA)

ExoMars (ESA)

Return

* Endorsed by the **International Planetary Data Alliance** in July 2012 – https://planetarydata.org/documents/steering-committee/ipda-endorsements-recommendations-and-actions

PDS4 – Some Principles for Agile Data Curation

Thank you!

This research was carried out by the Jet Propulsion Laboratory,
managed by the California Institute of Technology
under a contract with the National Aeronautics and Space Administration.

Adopted Standards

- Uses the OAIS ***Information Object*** concept exclusively and recursively
  - ***Data Object*** – *Digital, Physical, and Conceptual*
  - ***Representation Information*** – *Describes the Data Object*

- Best practices from the digital archive, object registry, metadata data registry, information architecture, data exchange reference models were used.
  - *ISO 14721:2003 - Open Archival Information System (OAIS) Reference Model - Provides a standard for information objects.*
  - *ISO/IEC 11179:3 Registry Metamodel and Basic Attributes specification - Adopted for the data dictionary schema.*
  - *Reference Architecture for Space Information Management (RASIM) - CCSDS 312-0.G-1 – Provides the overarching architectural principles.*
  - *W3C XML (Extensible Markup Language) - Rules for encoding documents electronically.*
  - *W3C XML schema - Type description language for XML documents.*
  - *Electronic Business XML (ebXML) federated registry/repository information model – Provides a standard to support federated registry/repository functions*
  - *RDF/RDFS/XML - RDF is a standard model for data interchange on the Web.*